The Gene of Scientific Success

Article *in* ACM Transactions on Knowledge Discovery from Data · May 2020

DOI: 10.1145/3385530

6 authors, including:

Xiangjie Kong
Zhejiang University of Technology
126 PUBLICATIONS **2,051** CITATIONS

SEE PROFILE

The Gene of Scientific Success

XIANGJIE KONG, Zhejiang University of Technology, China
JUN ZHANG, Dalian University of Technology, China
DA ZHANG, University of Miami, USA
YI BU, Peking University, China
YING DING, University of Texas at Austin, USA
FENG XIA*, Federation University Australia, Australia

This paper elaborates how to identify and evaluate causal factors to improve scientific impact. Currently, analyzing scientific impact can be beneficial to various academic activities including funding application, mentor recommendation, and discovering potential cooperators etc. It is universally acknowledged that high-impact scholars often have more opportunities to receive awards as an encouragement for their hard working. Therefore, scholars spend great efforts in making scientific achievements and improving scientific impact during their academic life. However, what are the determinate factors that control scholars' academic success? The answer to this question can help scholars conduct their research more efficiently. Under this consideration, our paper presents and analyzes the causal factors that are crucial for scholars' academic success. We first propose five major factors including article-centered factors, author-centered factors, venue-centered factors, institution-centered factors, and temporal factors. Then, we apply recent advanced machine learning algorithms and jackknife method to assess the importance of each causal factor. Our empirical results show that author-centered and article-centered factors have the highest relevancy to scholars' future success in the computer science area. Additionally, we discover an interesting phenomenon that the h-index of scholars within the same institution or university are actually very close to each other.

CCS Concepts: • **Information systems** → **Information systems applications**; • **Social and professional topics** → **Professional topics**;

Additional Key Words and Phrases: Scientific Impact, Academic Networks, Machine Learning, Feature Selection

*Corresponding author

The authors would like to thank Shenwei Zhang and Wenjie Kang for their helps with the experiments.
Authors' addresses: Xiangjie Kong, College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, 310023, Zhejiang, China, xjkong@acm.org; Jun Zhang, Graduate School of Education, Dalian University of Technology, Dalian, 116024, Liaoning, China, junzhang@dlut.edu.cn; Da Zhang, Department of Electrical and Computer Engineering, University of Miami, 5452 Coral Gables, Miami, FL, 33124, USA, zhang.1855@miami.edu; Yi Bu, Peking University, Department of Information Management, Beijing, 100871, China, buyipku@gmail.com; Ying Ding, University of Texas at Austin, School of Information, Austin, TX, 78701, USA, ying.ding@austin.utexas.edu; Feng Xia, School of Science, Engineering and Information Technology, Federation University Australia, Ballarat, VIC, 3353, Australia, f.xia@acm.org.

ACM Reference Format:
Xiangjie Kong, Jun Zhang, Da Zhang, Yi Bu, Ying Ding, and Feng Xia. 2020. The Gene of Scientific Success. 0, 0, Article 00 (2020), 20 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

The development of our society highly associates with scientists' diligent research comments. To recognize and support their contributions, a series of awards and research opportunities are given to these outstanding researchers. Recently, many researchers focus on how to determine the outstanding scholars and how to become a successful scholar, and many indicators from different aspects are proposed to quantify this scientific success [19, 25]. However, among these diverse indicators, which factor(s) is decisive and contributive to the academic success remains to be explored. Moreover, causal effects learning is a fundamental problem in machine learning with applications in various fields such as biology, economics, epidemiology, and computer science [31, 38]. Inspired by the above observations, our paper focuses on learning the causal effects that play vital roles in scholars' academic success.

Quantifying the scientific success of scholars has always been an interesting topic that attracts researchers with diverse backgrounds to study [13, 30, 37, 40]. The goal of science of success is to first understand the underlying mechanism and then discover a generative model to predict what the values of scientific success would be taken into account outside causal factors. Based on citation counts, a series of evaluation metrics have been proposed, such as journal impact factor [14], g-index [12], and h-index [16] etc. B. Van Houten [34] defines scientific impact as peer evaluation of scientific research academic works and other achievements, the importance of scientific impact depends on their research achievements being valued, recognized and cited by others. To measure the scientific impact, researchers have identified various controlling factors to capture the diverse characteristics of scholarly entities. Among these factors, citation counts have been regarded as the primary factor to evaluate the scientific impact for its simplicity and efficiency.

Besides the citation-based metrics, scientists also investigate this question from the perspective of network topologies. Initially, the importance of ranking algorithms, such as PageRank [27] and HITS algorithms [17], are designed for ranking web pages' importance. Recently, inspired by these importance ranking algorithms, researchers also widely utilize them to evaluate the scientific impact in academic networks [2, 11]. Considering the merits of both PageRank and HITS algorithms, Wang et al. [36] propose the MRCoRank method to measure the impact of scholarly entities in heterogeneous academic networks through mutual reinforcement.

Additionally, the development of social media enables scholars share their articles regularly on Twitter or Facebook. This information sharing effect is no more limited to academic social networks. It has spread out to many digital libraries and is prevalently used across social media platforms. Along with this trend, Altmetrics is proposed as another benchmark to measure the popularity of scholars and their publications by assessing their obtained social attentions. Meanwhile, researchers start utilizing the Altmetrics to quantify the researcher's scientific impact since it can capture the early impact promptly. For instance, Bornmann et al. [6] normalize publications' Twitter counts to measure the impact of research, and then use it for cross-field comparisons. Furthermore, lots of research methods explore the correlation between Altmetrics and citation counts based methods by statistically analyzing their interrelationship [9, 39].

The evaluation of scientific impact can shed light on diverse practical issues, such as awards or funding applications, job employments and advisor choosing [24]. Commonly,

Fig. 1. Illustration of a scholar's impact relevant factors.

successful scientific scholars obtain extra opportunity to acquire research resources, receive grant, and spread their research idea more extensively. Therefore, scholars aspire to improve their scientific impact continually. However, what factors have causal relationships with scientific success? Additionally, which factors are most appropriate to evaluate the scientific success? The above questions still remain unresolved. Therefore, in this paper, we conduct researches on first, identifying causal factors that contribute to the scientific success and then gauge the causality importance of these factors. In this paper, we take the most commonly used h-index as the metric for evaluating scholars' impact and mine the causal factors that lead to scholar's high h-index.

Due to privacy issues and technology limitations, the publications' information is easier to access compared to other data sources. It can also represent the corresponding scholars' academic abilities or contributions considerably. Generally, the title, keywords, authors, institutions, venue, pages, and published dates of a publication can be directly obtained. Based on these data, a lot of impact factors can be extracted and calculated as most of the current work does. While unlike previous work, our method does not focus on improving the evaluation metrics, instead we are aiming at discovering the causal factors that affect scholars' academic success from their publications. To answer this question, we categorized the impact factors into several categories as shown in Figure 1. They are article factor, author factor, venue factor, institution factor, and temporal factor. For each factor, we propose concrete and intuitive indicators to represent each scholar's academic characters. After that, by utilizing the machine learning algorithms and jackknife method, we explore the contribution of each factor on scholars' academic success.

Contribution. Our research mainly focuses on discovering causal factors of scholars' academic success. In general, we make the following contributions in this paper:

- **Novel features.** We present five potential causal factors taking the novel Gini Coefficient of institutions into account.

- **Causal detection.** Through utilizing the machine learning algorithms and jackknife method, we find that scholars' author-centered and article-centered factors are highly correlated with their academic success.
- **New insight.** Our findings provide researchers a novel and efficient method to improve their scientific impact.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 identifies the proposed scientific impact factors. Section 4 verifies the causal factors, assess their significance and testify them over big scholar dataset. Then, we conclude our work in Section 5.

2 RELATED WORK

The scientific impact has been studied for decades by researchers from a variety of disciplines. For a long time, citation counts have been widely applied to measure the scientific impact. Along with this tendency, researchers have proposed various citation-based metrics. With the developments of academic networks, scholars also look into the scientific impact problem from the angle of network importance. These indicators can be employed for both impact evaluation and future academic success predictions. In this section, we will introduce the related work from the above-mentioned aspects respectively.

The citation count was first utilized to quantify the impact of journals. From then on, researchers have proposed a variety of citation-based methods to measure the scientific impact [33], such as the h-index [16] and g-index [12]. Some scholars claim that citations should not be regarded as equal [5]. Another example that should be mentioned here is: Both A and B cite C. Previously the citations from A and B are regarded as equal, but if A is from a highly-cited paper while B is not, the citations should be differentiated. This is similar to PageRank-related ideas. A series of approaches for distinguishing the importance of citations have been proposed [33]. These methods all applied the citation counts as an important part of the evaluation metrics, while they all make some improvements since simply relying on citation counts is unilateral for impact evaluation [35].

Besides utilizing citation to quantify the scientific impact, scholarly networks are now frequently applied to study such problems since the networks contain various types of entities and relationships [23]. The PageRank and HITS algorithms are widely used for measuring the scientific impact in academic networks. On the basis of these two algorithms, a series of network-based evaluation metrics have been proposed [42]. Considering the effect of different academic network structures, scholars apply the modified importance ranking algorithms to evaluate the impact of different scholarly entities [3]. Other than considering network topologies, some researchers also discover novel features and relationships to evaluate the scientific impact. Wang et al. [36] rank the impact of scholarly entities by exploring the text features in heterogeneous academic networks. Due to the evolvement nature of academic networks, some studies also consider the dynamics of citations and the new emergence of new entities or relationships to evaluate the scientific impact [4, 43].

Other than using citation-based and network-based features to evaluate the scientific impact, scholars also try to explore the relevant factors which are very crucial for the future academic performance and predict the future impact [15, 26]. Wang et al. [35] verify the effectiveness of early citations in predicting the potential citations of articles [7, 18]. Stegehuis et al. [32] utilize two significant factors, namely historical citation information and journal impact factor to predict papers citation distribution.

The prediction of scholars' future influence, h-index, and future citations are all within the scope of future impact prediction [8, 35]. Acuna et al. [1] apply the number of papers, h-index, and academic ages of a scholar to predict his/her future impact. The linear regression method is utilized to predict the future impact of outstanding scholars from mathematics, physics and biology research area. And they found that the academic ages of scholars actually play a significant role in predicting scientific impact [22]. Additionally, Dong et al. [10] study the question of which paper can increase scholar's h-index through the linear regression method. They discover that among six factors, the topic and venue are very crucial for the predictions.

Scientific impact prediction with causal inference is a recently emerging research field. Unlike previous methods [1, 8, 35], causal inference based methods first identify potential causal factors and then use them to guide the scientific impact prediction. Additionally, previous methods only consider a single perspective for assessing the scientific impact while neglecting analyzing and ranking the importance of each causal factor.

3 CAUSAL FACTOR IDENTIFICATION

Researchers have studied the problem of scientific impact for decades and propose a variety of impact factors. However, there is no formal definition for scientific impact and no commonly accepted standard for scientific impact evaluation up to now. Among these previously studied impact factors, which factors are most relevant to scholars' academic success? Discovering the answer to it can help researchers carry out their research more efficiently. In this section, we will introduce several novel impact factors, organize the existing factors, and classify them into different categories.

3.1 Article-centered Factors

Generally, most previous work prefers using the citation counts and the number of articles to quantify the scientific impact. While beyond these two indicators, there exist diverse article-based factors that affect the dynamics of scientific impact. To discover the representative features for articles, we first analyze the elements related to article-centered factors.

Citation counts ($Cits$), and the number of publications (Num_{pub}) are the basis of article-based factors. The average citations for each scholar ($Ave_{ci}^{a_i}$), the highest citations ($Hi_{ci}^{a_i}$), the lowest citations ($Lo_{ci}^{a_i}$) can be directly obtained through the values of their total $Cites$ and (Num_{pub}). Moreover, the quality of an article depends not only on its content, but also on its topic popularity. For instance, previously, a wide variety of data cannot be acquired and processed due to the technical limitations. While with the developments of data processing technologies and advancement of big data era, paper related to big data topics receive more attention recently. Consequently, the topics of an article also can affect its influence. In order to capture this character, we propose the article's topic popular degree (ATP), which can be calculated according to the following equation.

$$ATP(p_i) = \frac{\sum_{w=1}^{m} Num(w)^{p_i}}{\sum_{i=1}^{n} Num(i)} \tag{1}$$

where p_i represents the paper, w is the keyword of paper p_i, $Num(w)$ is the number of w, m is the number of keywords in paper p_i, $Num(i)$ is the number of the keywords of papers, and n is the total number of publications.

Besides the above mentioned citation-based factors, the qualities of references also need to be considered when measuring the scientific impact. Generally, every scholar has a list of publications, and each publication has a series of references. Citations can be deemed as

academic acknowledgments from other researchers. Similarly, the authors of an article also are enlightened by its references. Therefore, references can affect the quality of an article. Primarily, the highest (Hi_{ci}^{ref}), the average (Ave_{ci}^{ref}), the lowest citation counts (Lo_{ci}^{ref}), and the average number of references (Ave_{num}^{ref}) are the most direct measurements to quantify the qualities of references. Beyond the citations, the impact of references' venues is also utilized to evaluate the impact of references since many researchers tend to cite articles from high impact venues regardless of the relevance between articles.

To measure the relevance between articles (Rel_{ref}), we first solve this problem from the angle of authors. According to each author's publications, their research areas can be represented by exacting articles' keywords. Therefore, we utilize the differences among authors' keywords to calculate the relevance between articles and references. The information entropy is applied to quantify it, and the calculation formula is as follows:

$$Rel_{ref}^{p \to q} = -\sum_{i=1}^{r} W_i \log_2 (W_i) \tag{2}$$

where $Rel_{ref}^{p \to q}$ represents the relevance between article q and its reference p, W_i is word's frequency in article q and p's keywords' information, and r is words' total counts.

Furthermore, the relevance between the articles and their reference also needs to be considered. Due to unavailability of articles' full texts, we use the cosine similarity to measure the relevance between the articles' and their references' titles and keywords. For each article and its reference, we extract the sequence of words $(m_1, m_2, m_3, ..., m_n)$ from their titles and keywords. Then a vector can be obtained based on the above sequence for each paper. According to these vectors, the relevance between an article and its reference can be calculated as follows:

$$Sim(p_1, p_2) = \frac{\sum_{i=1}^{n} (V_{p_1,i} * V_{p_2,i})}{\sqrt{\sum_{i=1}^{n} V_{p_1,i}^2} * \sqrt{\sum_{i=1}^{n} V_{p_2,i}^2}} \tag{3}$$

where $Sim(p_1, p_2)$ represents relevance between paper p_1 and p_2, V_{p_1} is vector of p_1, and V_{p_2} is p_2's vector.

3.2 Venue-centered Factors

Besides citation-based metric, PageRank can also be used to measure the qualities of the venues, which reflects the scientific success of an author. To gauge the importance of venues, the PageRank values ($PR(v_i)$) in the paper-venue network are first calculated. Then, the average citations of papers published in the venues ($Ave_{ci}^{v_i}$) is used to measure the quality of venues. Furthermore, with the aids of the concept of scholar's h-index, we calculate the h-index of venue ($h(v_i)$). Specifically, the definition of the h-index of a venue is similar to the original calculation procedure of h-index, and the h-index value of a venue equals to h that at least h papers in the venue have h citations.

$$PR(v_i) = \sum_{j=1}^{n} Ave_{cj}^{v_i} \tag{4}$$

3.3 Author-centered Factors

Aside from article-centered factors, the factors that represent scholars' attributes are vital to their impact as well. Each scholar's h-index (h_{a_i}) and PageRank value (PR_{a_i}) in collaboration network are intuitive factors to indicate a scholar's impact. Meanwhile, the journal impact

factor (JIF) can be calculated based on them, and is widely applied to measure the impact of journals for its simplicity. According to the concept of JIF, the author's impact factor (AIF) is proposed. Similarly, the AIF of a scholar in year t is scholar's Ave_{ci} in $\triangle t$ years before year t. Besides the citation counts, the sum of PageRank scores of scholars' papers PR_{pub} in citation network also can indicate their importance.

Other than these two factors, scholars have proposed several well-known factors to quantify the dynamics of scholars' impact. The Q value is widely applied to reveal the mutual reinforce process of scholars' impact on their papers [30] and is stable during scientists' whole academic careers. The calculation formula of Q value is as follows:

$$Q(a_i) = e^{\langle logc_{i\alpha} \rangle} - \mu_p \tag{5}$$

where $Q(a_i)$ represents scholar Q value, $\langle logc_{i\alpha} \rangle$ is a_i's average citations in logarithmic way, α is a_i's α-th article, and μ_p is the average potential influence of articles.

While in each scholar's academic career, they will encounter a variety of researchers from different disciplines. Scholars will benefit from the academic exchanges and discussions with other researchers, and furthermore improve their own scientific impact. As a consequence, the capacities of coauthors also can affect the qualities of their articles and scholars' impact in the meantime. To capture coauthors' influence, several factors are proposed. Typically, the h-index of coauthors represents their abilities. Based on it, a series of factors can be easily obtained. The max ($hmax_{co}^{a_i}$) and average values ($have_{co}^{a_i}$) of scholar a_is' coauthors can be directly acquired through each author's h-index. Then we use the differentials between a_i's h-index and $hmax_{co}^{a_i}$ ($hdif_{a_i}$) to represent the distance between influential coauthors and a_i.

Besides using the h-index to quantify coauthors' academic capacities, we then consider the effects of coauthors' diverse research backgrounds on scholars' impact. Since co-operations among researchers are getting more and more frequently, the integration of scholars from multi-disciplines also has the positive influence on promoting the developments of science and technologies. To measure the range of coauthors' disciplines, we apply the theory of entropy. The detail information on scholars' specific disciplines and institutions can be obtained from the dataset. For each scholar, we quantify the diversity of his or her coauthors ($Div(a_i)$) by utilizing the theory of entropy. The diversity is computed according to the following equation.

$$Div(a_i)_{inst} = -\sum_{m=1}^{r} w_m \log_2 (w_m) \tag{6}$$

$$Div(a_i)_{key} = -\sum_{\rho=1}^{q} k_\rho \log_2 (k_\rho) \tag{7}$$

$$Div(a_i) = Div(a_i)_{inst} + Div(a_i)_{key} \tag{8}$$

where $Div(a_i)_{inst}$ and $Div(a_i)_{key}$ represent author a_i's diversity of cooperators' institutions and their papers' keywords, and $Div(a_i)$ indicates a_i's overall cooperators' diversities. w_m is word m's frequency in the overall a_i's cooperators' institutions' information, and r is word m's total counts in Eq. (9). k_ρ is word ρ's frequency in all a_i's cooperators' papers' keywords, and q is the total number of word ρ.

3.4 Institution-centered Factors

The effects of institutions on scholars' impact also need to be considered since research funding or policy issues can significantly influence researchers' progress on their studies. Meanwhile, scholars' academic achievements also can be affected by the capacities of their

colleagues because they may frequently share research ideas and techniques. Generally, we explore the effects of institutions from two major aspects: scholars' academic environments and the economic factors.

We measure the academic environments from the perspective of colleagues. When conducting researches, people tend to exchange idea with their co-authors or colleagues. Additionally, researchers are also affected by the influencing group or individuals in their institution. This influence is usually defined as peer pressure (or social pressure). Taking this peer pressure influence into account, we try to identify the relevance of peer pressure on scholars' academic performance. In other words, is there an actual relationship between them? To answer the above questions, we proposed several factors to reveal the correlation between scholars' academic success and their colleagues. Initially, we gauge the research capacities of scholars' colleagues. For each scholar, his or her colleague's h-index (h_{col}), number of publications (Num_{pub}^{col}), citation counts ($Cits_{col}$), and PageRank score (PR_{col}) can be calculated over the dataset.

Furthermore, we employ the concept of Gini coefficient from the economic field to describe academic reputation of an institution. The Gini coefficient originally utilizes the definition of Lorenz global curves to compute the distribution of income in economic field. Its value ranges from 0 to 1. The bigger the value is, the more economic inequality is. In our paper, we quantify the Gini coefficient of institutions using the values of scholar's h-index, citations, and the number of papers. The Gini coefficient of an institution can be calculated as follows:

$$G(i) = 1 - \frac{1}{n}(2\sum_{m=1}^{n-1} P_m + 1) \tag{9}$$

Here, $G(i)$ represents the Gini coefficient value of institution i and n is the number of research groups within the institution i. For a research group, m indicates the group index among n groups. Also, P_m is the proportion of the sum of group m in the whole values of institution i. Therefore, according to the values of scholar's h-index, citations, and the number of papers, the reputation of each institution are calculated using three Gini coefficient values which are $G(i)^h$, $G(i)^{Cit}$, and $G(i)^{pub}$.

3.5 Temporal Factors

Previous studies have verified the effect of temporal dynamics on scientific impact, such as predicting academic rising stars. For young researchers, they may have a fast growth stage after starting the academic career. The performance during this period is very crucial for their future academic success. We propose two temporal factors to capture this phenomenon. The first one is the academic ages (Num_{years}), which are the years since scholars publish their first academic papers. Another factor is scholars' dynamics of h-index during $\triangle t$ years. In this paper, we set the $\triangle t = 3, 5, 7$, and then calculate the difference ($Hindex\text{-}dif$) between the predicted time and $\triangle t$ years ago.

$$\rho = \frac{cov(X,Y)}{\sigma X \sigma Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{10}$$

where cov is the covariance between two groups of results, and σ indicates their standard deviation. Its value ranges from -1 to 1 with correlation varying from the most negative to the most positive.

From the above-mentioned factors, we try to list the relevant indicators of scholars' academic success as comprehensive possible. These indicators are categorized into five major

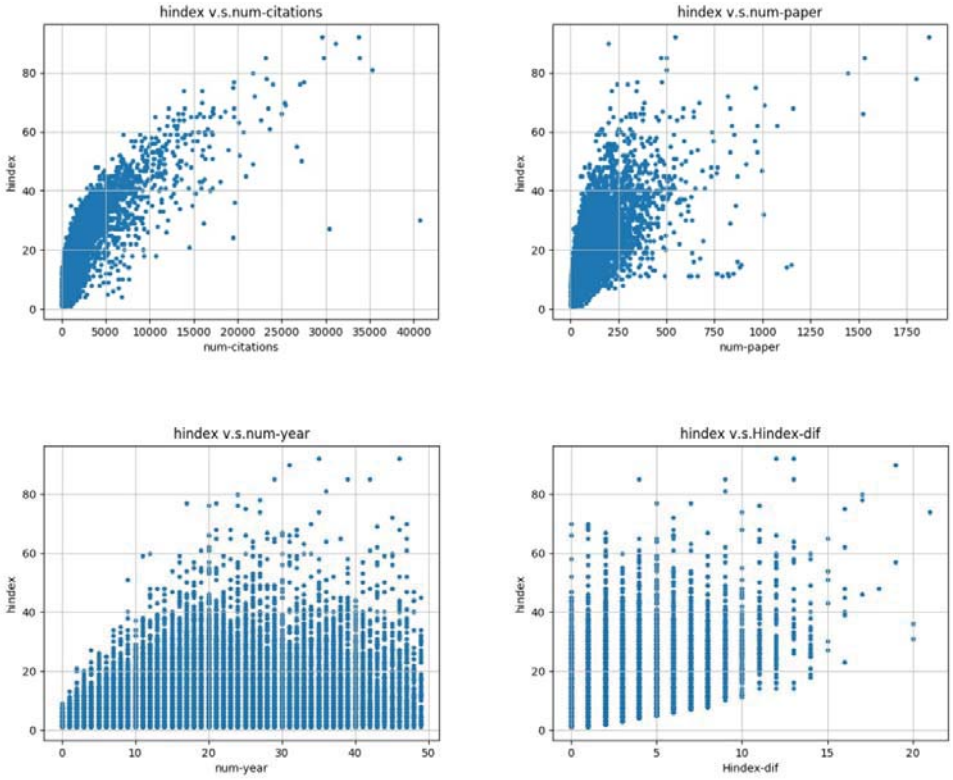

Fig. 2. The linear regression phenomena between the h-index and the four most relevant factors.

categories, which are article-centered factors, author-centered factors, venue-centered factors, institution-centered factors, and temporal factors. These factors are described in Table 1. Meanwhile, we primarily investigate the correlation between these factors and scholar's h-index through the most direct way. The Pearson Correlation Coefficient is applied to measure the relevance between two ranking results. The calculation procedure of Pearson Correlation Coefficient is shown as follows:

According to Table 1, we can see that the author-centered and article-centered factors are the most correlative factors among all the proposed indicators following by the temporal-centered factors. To further present the linear correlation phenomenon between the h-index and the other factors, we then depict the results of the four most relevant factors. As shown in Figure 2, scholars' number of citations and publications, academic years, and the differentials of h-index in $\triangle t$ years are highly correlate with scholars' future h-index. With the increase of academic age, h-index keeps increase until academic age is 15, after that keeps steady.

It is not difficult to understand the high relevance of $Cites$, Num_{pub}, and h_{a_i}. While the venue-centered and institution-centered factors seem to be negatively correlated with scholar's h-index. However, this table cannot accurately depict the effectiveness of these factors on predicting the future academic success of scholars since the factors may reveal the same phenomenon together. From this table, it can only show the linear correlation between

Table 1 Causal Factor Descriptions and Correlations.

	Feature	Description	Correlation
Article	$Cits$	The citation counts of scholars.	0.7629
	Num_{pub}	The number of publications of scholars.	0.7782
	Ave_{ci}	The average citations of each scholar.	0.2772
	Hi_{ci}	The highest citations of each scholar.	0.2349
	Lo_{ci}	The lowest citations of each scholar.	0.2067
	ATP	The article's topic popular degree.	0.0134
	Hi_{ci}^{ref}	The highest citations of references.	0.1648
	Ave_{ci}^{ref}	The average citations of references.	0.1439
	Lo_{ci}^{ref}	The lowest citations of references.	0.0648
	Ave_{num}^{ref}	The average number of references.	0.2496
	Rel_{ref}	The relevance between articles.	0.0174
	$Sim(p_1, p_2)$	The cosine similarity between articles.	0.1437
Venue	$PR(v_i)$	Venues' PageRank values in the paper-venue network.	0.2146
	$Ave_{ci}^{v_i}$	The average citations of papers published in venues.	0.1924
	$h(v_i)$	The h-index of venues.	0.2081
Author	$h(a_i)$	Each scholar's h-index value.	0.9782
	PR_{a_i}	Each scholar's PageRank value in co-author network.	0.6274
	AIF	The author impact factor.	0.3826
	$Qvalue$	The author's Q value.	0.5394
	$hmax_{co}^{a_i}$	The max h-index value of scholar's coauthors.	0.8253
	$Num_{co}^{a_i}$	The number of scholar's coauthors.	0.426
	$have_{co}^{a_i}$	The average h-index value of scholar's coauthors.	0.482
	$hlo_{co}^{a_i}$	The lowest h-index value of scholar's coauthors.	0.275
	$hdif_{a_i}$	The differentials between the max and the lowest h-index value of scholar's coauthors.	0.538
	$Div(a_i)$	The diversity of coauthors.	0.1743
Institution	h_{col}	The h-index of scholars's colleague.	0.2947
	Num_{pub}^{col}	The number of publications of scholars's colleague.	0.1368
	$Cits_{col}$	The citation counts of scholars's colleague.	0.1937
	PR_{col}	The PageRank score of scholars's colleague.	0.0264
	$G(i)^h$	The Gini coefficient on h-index of institution.	0.0937
	$G(i)^{Cit}$	The Gini coefficient on citation counts of institution.	0.0153
	$G(i)^{pub}$	The Gini coefficient on number of publications of institution.	0.1632
	GDP	The GDP value of the institution's country.	0.1937
Temporal	Num_{years}	Scholar's academic ages.	0.5863
	$Hindex\text{-}dif$	The difference between scholar's h-index and $\triangle t$ years ago.	0.6248

them, and their performance in predicting scholar's h-index are investigated in the next section.

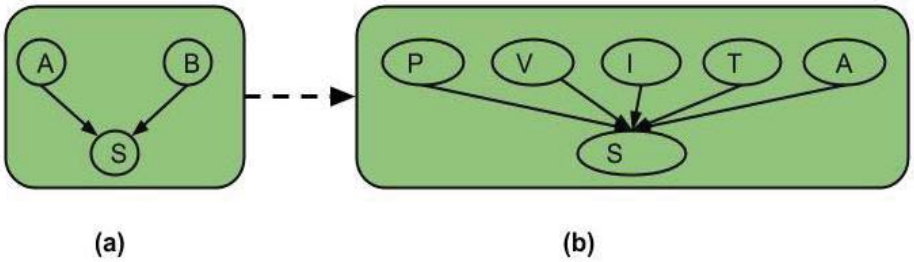

Fig. 3. V-structure causal inference model

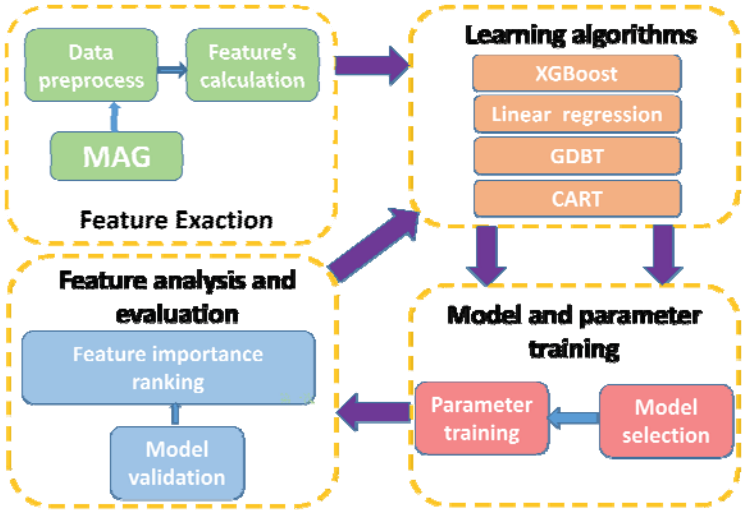

Fig. 4. Causal inference framework

4 CAUSAL FACTOR VERIFICATION

In this section, we explore the performance of the above-mentioned factors on predicting scholar's h-index. In order to investigate their effectiveness, we use advanced machine learning techniques. Among them, we apply the XGboost, linear regression, gradient boosting decision trees, and classification and regression tree separately on the dataset. Then, by comparing the performance, we find the most appropriate machine learning method.

4.1 Structured Causal Model (SCM) Construction

The core of the structural theory of causation lies a "structural causation model (SCM) [21, 29, 41]". Therefore, in our paper, we first present a simple causal model as shown in Fig. 3. Here, S is a collider: arrows 'collide' at S. the path $A \rightarrow S \leftarrow B$ is blocked. In other words, A is not associated with B through S. Given a collider S. the causal factors are independent with each other. We call this structure a V-structure [20]. In this V-structure, A and B represent parents of S. However, in our paper, scientific success S can be determined and

influenced by multiple causal factors. Therefore, we expand in Fig. 3(a) as a heterogeneous structured causal model as shown in 3(b) formulated as the following equation.

$$P(S) = P(S|F_1) \dots P(S|F_n) \qquad \text{Causal Factor Discovery} \qquad (11)$$

4.2 Causal Effects Prediction

A SCM model S, consisting of two sets of variables, X and Y ,and a set F of functions that determine how values are assigned to each variable $X_i \in X$. Here, we assume that, given the prediction output Y, the function f represents the effect Y as a function of the direct causes X and marginal loss ϵ with learning parameters θ_1. After we identify factors that contribute to our scientific success, we need to measure the significance for each causal factor. From the observational big scholarly dataset, we apply advanced machine learning techniques to discover the causal relationships and how various causal factors, including *Author, Paper, Venue, Institution and Temporal* facilitates understanding the scientific success.

$$S = F(X, Y, \epsilon; \theta_1) \qquad \text{Loss Function} \qquad (12)$$

In this section, we apply the following four advanced machine learning techniques to estimate the causal effects.

XGBoost: XGBoost is a scalable end-to-end tree boosting system and is faster than the most current widely used methods. The idea of the boosting algorithm is to integrate many weak classifiers together to form a strong classifier, and XGBoost is a lifting tree model which integrates many CART regression tree models to form a strong classifier. Its tree boosting mainly consists two parts, which are the regularized learning objective and the gradient tree boosting process.

Linear Regression (LR): Regression analysis is widely used for prediction and forecasting, and it can also be used to find out which among all independent variables are related with the dependent variable. Linear regression requires that the model is linear in regression parameters. The predictor function is utilized to model the data, and the data can be used to estimate the unknown parameters. Linear Regression is fast in modeling and runs fast in the case of large amounts of data.

Gradient Boosting Decision Trees (GBDT): GBDT is an iterative decision tree algorithm, which includes many decision trees and the final result equals to the sum of all the trees' decisions. The core of GBDT is that every tree learns the residual of the sum of all previous tree conclusions, which is the sum of the real values after adding the predicted values. It can discover a variety of distinct features and their combinations.

Classification and Regression Trees (CART): It can be used to create a classification tree or a regression tree. When CART is used as a classification tree, the feature attributes can be continuous or discrete, and a CART classification tree uses Gini index in node splitting. When CART is used as regression tree, observation attributes are required to be continuous type. Because the least absolute deviation (LAD) or least square deviation (LSD) method is usually used when selecting feature attributes by node splitting, the feature attributes are also continuous type. In our paper, we apply it as a regression tree to predict scholar's future impact based on the input variables.

4.3 Dataset

In this paper, we use two datasets of different disciplines. One is the sub-dataset extracted from the Microsoft Academic Graph (MAG). The MAG dataset contains detailed paper information including *title*, *keywords*, *authors*, *institutions*, *venues*, *publication date*, and *citations* from 27 macro-areas and 306 sub-areas. The whole dataset includes over 35 million papers, 38 million authors, and more than 324 million citation relationships. We use a sub-dataset includes 79,321 scholar profiles and 105,123 articles focusing on computer science domain with complete academic careers.

The other is a subset of American Physical Society (APS). The APS dataset contains physics paper information of *title*, *authors*, *institutions*, *venues*, *publication date*, and *citations*. The whole dataset includes 540,232 papers, 394,801 authors, and more than 6 million citation relationships of 12 APS journals. We use a sub-dataset including PRC and PRE papers, 80,360 scholar profiles and 98,011 articles in total.

4.4 Evaluation Metrics

In order to evaluate the performance of different learning algorithms and factors, we adopt four typical metrics including MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), MSE (Mean Squared Error), ACC (Accuracy), and R^2. Given the true value y, and the predictive value $\hat{y}$, the above-mentioned evaluation metrics can be calculated as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y| \tag{13}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}|\frac{y - \hat{y}}{y}| \tag{14}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y} - y|^2 \tag{15}$$

$$ACC = \frac{1}{n}\sum_{i=1}^{n} I(f(y_i) = \hat{y}) \tag{16}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}|y - \hat{y}|^2}{\sum_{i=1}^{n}|y - \bar{y}|^2} \tag{17}$$

4.5 Validation by experiments

With the learning algorithms and factors we introduced above, scholar's h-index can be predicted. We use the previous $\triangle t$ years' information for training, and the real data in 2015 (MAG) or 2013 (APS) to validate. On the train set, we perform a 5-folds cross-validation to tune the hyperparameter of models. For XGBoost and GBDT, the main hyperparameters we have tuned include the learning rate, maximum depth of trees, sub-sample rate, sub-feature rate and the regularization coefficient. For CART, the main hyperparameters we have tuned include the learning rate, maximum depth of trees. For LR, the main hyper parameters we have tuned include the learning rate and the regularization coefficient. All hyper parameters

are tuned by grid search on the parameter space. The results are illustrated from the aspect of MAE, MAPE, MSE, ACC, and R^2.

Table 2 shows the predictive performances of different methods on the evaluation metrics mentioned above on MAG dataset. MSE, MAE, and MAPE are used to compare the predictive results and the true values. In the table, R^2 indicates the correlation between the predictive results and the true values and ACC indicates the accuracy. Hence, the better prediction performance can be inferred by their values. It is obvious that the performance of XGBoost is the best among all the methods using different time periods because it outperforms other methods on 4 of 5 metrics, which gets the smallest MAPE and MES and highest Acc and R^2 score in three groups experiments. While for different $\triangle t$ values, there exist various prediction results. The performance of $\triangle t$=7 achieves the best score, and the results by $\triangle t$=10 are the worst. However, there only exists a slight difference between the results of $\triangle t$=5 and $\triangle t$=7. In the following parts, we analyze the results in $\triangle t$=7 on the MAG dataset.

Table 2 The Performance of Difference Learning Algorithms on MAG.

		MAE	MAPE	MSE	ACC	R^2
$\triangle t$=5	XGBoost	0.73	0.07	1.09	0.86	0.99
	LR	0.82	0.10	1.18	0.80	0.92
	GBDT	0.69	0.08	1.16	0.84	0.95
	CART	0.96	0.18	2.30	0.79	0.81
$\triangle t$=7	XGBoost	0.79	0.07	1.19	0.86	0.99
	LR	0.83	0.11	1.30	0.79	0.90
	GBDT	0.73	0.09	1.25	0.85	0.94
	CART	0.98	0.20	2.49	0.78	0.80
$\triangle t$=10	XGBoost	0.81	0.09	1.27	0.83	0.91
	LR	0.84	0.13	1.43	0.73	0.86
	GBDT	0.74	0.10	1.32	0.81	0.84
	CART	0.99	0.29	2.68	0.74	0.63

The predictive performances of these methods on the APS dataset are shown in Table 3. In the table, we observed that the overall prediction performances on APS are better than on MAG, which has smaller errors and higher fitting degree (R^2). We noticed that the performance of XGBoost is also the best. Different from the results on MAG, all methods perform better on the period $\triangle t = 10$ than other groups. In the following parts, we analyze the results in $\triangle t$=10 on the APS dataset.

4.6 Causal Factor Evaluation

The above analyses verify the causal relationships between various factors and the overall h-index results. However, the contribution and importance of factors still need to be explored. To solve this question, we first apply the "jackknife" method [28] to verify the function of each group's factors separately. The "jackknife" method includes two phases: *Adding* and *Removing*. During *Adding* phase, we use one group of factors each time to predict the result. During *Removing* phase, we remove a group of factors and train the model with the rest factors. After these two phases, factor's individual contribution to the overall prediction task can be explored.

Table 3 The Performance of Difference Learning Algorithms on APS.

		MAE	MAPE	MSE	ACC	R^2
$\triangle t$=5	XGBoost	0.54	0.06	0.62	0.81	0.97
	LR	0.57	0.08	0.65	0.80	0.96
	GBDT	0.55	0.06	0.63	0.80	0.96
	CART	0.56	0.09	0.73	0.78	0.95
$\triangle t$=7	XGBoost	0.51	0.07	0.55	0.82	0.97
	LR	0.53	0.10	0.55	0.80	0.96
	GBDT	0.51	0.07	0.56	0.81	0.97
	CART	0.54	0.08	0.70	0.80	0.96
$\triangle t$=10	XGBoost	0.45	0.06	0.46	0.85	0.98
	LR	0.47	0.07	0.48	0.83	0.97
	GBDT	0.45	0.08	0.47	0.84	0.97
	CART	0.46	0.07	0.60	0.80	0.96

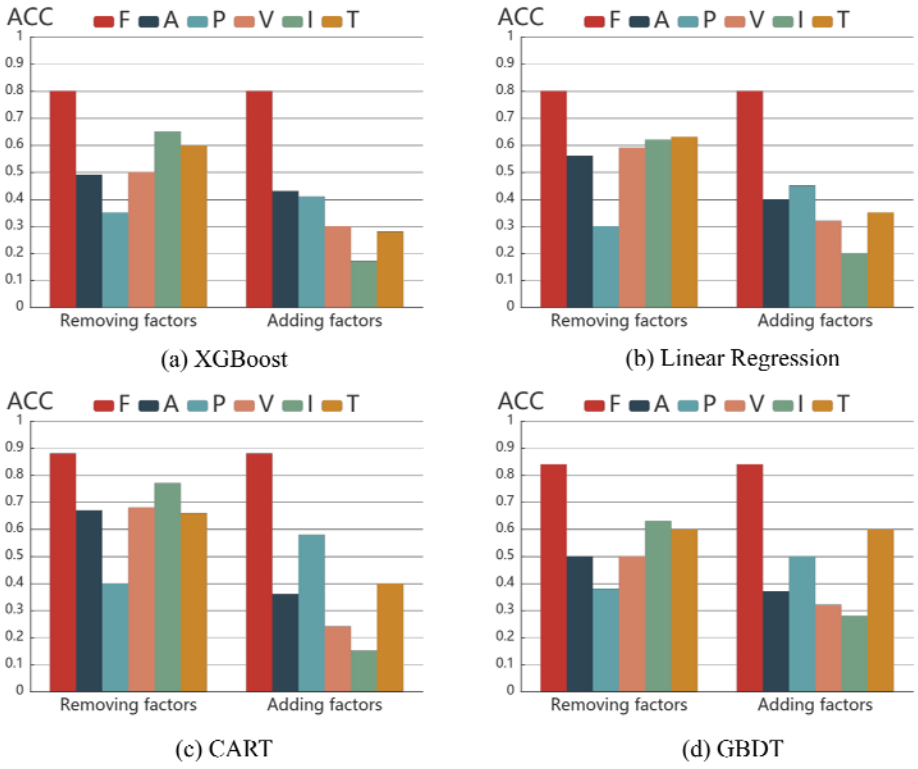

Fig. 5. Factor contribution analysis on MAG. Four models trained with only or without the denoted factors. F: full feature set; A: Author factors; P: Paper factors; V: Venue factors; I: Institution factors; T: Temporal factors.

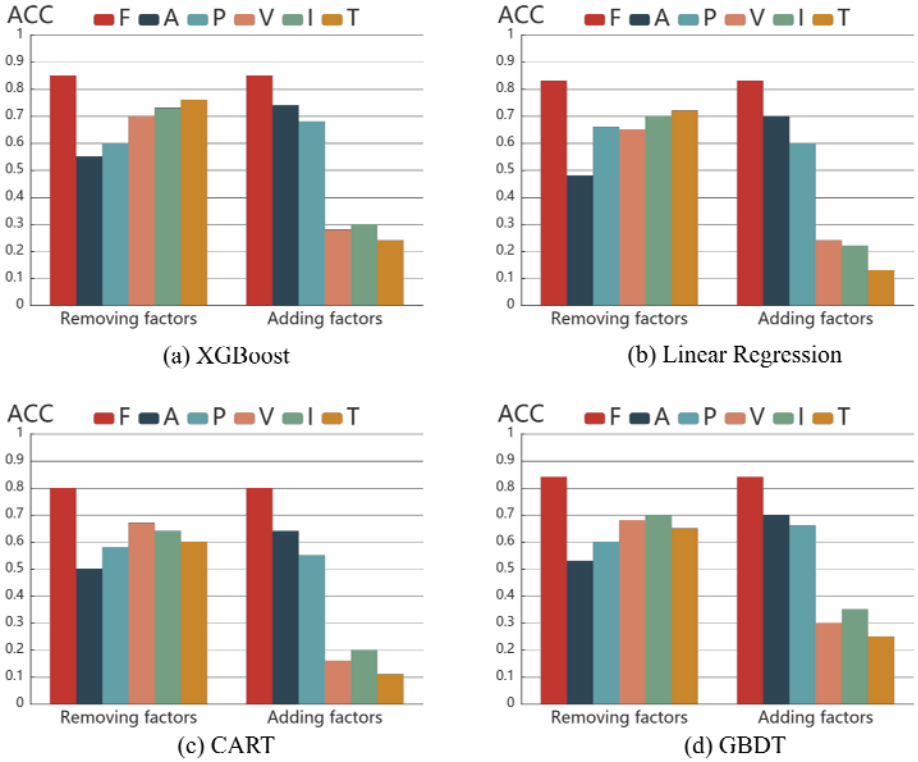

Fig. 6. Factor contribution analysis on APS. Four models trained with only or without the denoted factors. F: full feature set; A: Author factors; P: Paper factors; V: Venue factors; I: Institution factors; T: Temporal factors.

As shown in Figure 5, in the experiments on MAG dataset, the drop in ACC values by the removal of article-centered factors in the four methods demonstrate that they are of great significance in predicting the h-index. On the contrary, when removing other types of factors, the decline of the predictive performance is not so obvious. This fact can reveal the importance of article-centered factors on predicting scholars' future success. For adding factors, article-centered factors still show their important roles in predicting the future scientific impact. Moreover, author-centered and temporal-centered factors also show their effectiveness for scholar's h-index prediction.

The same analyses are performed on APS dataset, as shown in Figure 6. Different from the results on MAG, the drop in ACC values by the removal of author-centered factors greatly influences the predicting of h-index, which indicates that the author-centered factors are of great significance in the APS dataset. Same as previous experiments, when removing other types of factors, the decline of the predictive performance is not so obvious. For adding factors, author-centered factors still show their important roles. Moreover, article-centered factors also show their effectiveness for scholar's h-index prediction in APS dataset, but other features have no obvious effects, which differs from experiment results on MAG.

Furthermore, we analyzed the feature importance given by XGBoost of results on MAG and APS dataset. In XGBoost, feature importance can be calculated as times that a feature

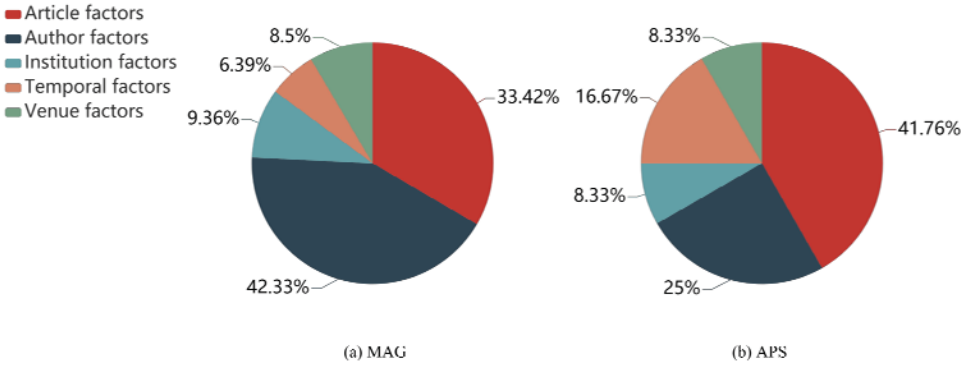

Fig. 7. The importance score of different factors.

has been used to divide samples on leaves of trees in the model. The more frequently a feature has been used, the more important it is to the model. As shown in Figure 7, in both datasets, the top influential factors are still the same with the above conclusions, where the article-centered factors are still the most important for predicting the future academic success in MAG, which takes 41.67% importance scores; and the author-centered factors are importance in APS, which takes 42.33% importance scores.

In addition, we also analyze the Gini Coefficient of different institutions. Gini Coefficient smaller than 0.2 indicates the institutions are absolute equal. Gini Coefficient from 0.2 to 0.3 indicates the institutions are relatively equal. Gini Coefficient from 0.3 to 0.4 indicates the institutions are relatively rational. Finally, Gini Coefficient greater than 0.4 indicates the great disparity between institutions.

To have a comprehensive Gini coefficient of an institution, we first rank the institutions according to the number of scholars. Then according to the ranking list, the Gini coefficient on citations, number of publications, and h-index can be obtained. We show the Gini coefficient of top 5%, 10%, 20%, and the last 10%. As shown in Table 4 and Table 5, there exist some interesting phenomena. For top 5% ranking institutions, both in MAG and APS, their Gini Coefficient values are under 0.2, which indicate that scholars' h-index is very close to their colleagues in the same institution. In top 10% ranking institutions, except for citations, the Gini Coefficients for the number of papers and h-index are under 0.2, which still show the equality of scholars on these two aspects. While for institutions in top 20% and last 10%, their Gini Coefficient for the number of papers and citation exceed 0.2. It is apparent that there exist some differences in the number of papers and citation of researchers in these institutions. However, the h-index level of scholars in all the institutions mentioned above is very similar to their colleagues. This phenomenon shows that scholars in the same institution are birds of s feather flock together. And this phenomenon is the same in computer science and physics. The reason behind this is that the scholarly communication among them is very convenient and frequent, and they can directly feel the peer pressure from their colleagues to some extent. Therefore, scholars are trying to keep up with their colleagues in academic research, and their overall scientific impacts are quite similar to each other. Also, when providing faculty positions for researchers, there may exist standard hiring requirements for the same institution. As a consequence, the scholars in the same institution are at the same academic level.

Table 4 The average Gini Coefficients of top ranking institutions on MAG.

	Number of papers	Citation	h-index
Top 5%	0.102418	0.161101	0.042667
Top 10%	0.191524	0.277041	0.091791
Top 20%	0.230564	0.327122	0.121142
Last 10%	0.351485	0.452952	0.200423

Table 5 The average Gini Coefficients of top ranking institutions on APS.

	Number of papers	Citation	h-index
Top 5%	0.011092	0.070037	0.015215
Top 10%	0.169559	0.205179	0.145778
Top 20%	0.216014	0.249943	0.184754
Last 10%	0.266891	0.297307	0.228714

5 CONCLUSION AND FUTURE WORK

In this paper, we aim at discovering the causal factors that play crucial roles in predicting the scholars' academic success. To solve this issue, we first propose five potential causal factors, which are the article-centered factors, author-centered factors, venue-centered factors, institution-centered factors, and temporal factors. Then by utilizing the state of the art machine learning algorithms, we find that the article and author-centered factors are most significant causal factors for forecasting scholars' future success.

Furthermore, we analyze each factor's contribution by using the "jackknife" method and grading factors during the predicting process. The results further demonstrate the importance of article and author-centered factors. We further analyze the specific importance ranking of these five groups of factors used in our experiments. After this process, we find that, in the MAG dataset, the article-centered factors have 41.47% importance, the author-centered factors are in 25% importance, the temporal-centered factors are in 16.67% importance, and the venue and institution-centered factors are in 8.33% importance, while in the APS dataset, the article-centered factors have 33.42% importance, the author-centered factors are in 42.33% importance, the temporal-centered factors are in 6.39% importance, and the venue-centered factors have 8.50% importance , and institution-centered factors are in 9.36% importance. Meanwhile, we also find that the h-index of scholars in the same institutions tend to be very close to each other.

In the future, we plan to identify more factors and conduct our experiments on other datasets from various disciplines to demonstrate the validity of our work.

REFERENCES

[1] Daniel E. Acuna, Stefano Allesina, and Konrad P. Kording. 2012. Future impact: Predicting scientific success. *Nature* 489, 7415 (2012).

[2] Tehmina Amjad, Ali Daud, Dunren Che, and Atia Akram. 2016. MuICE: Mutual Influence and Citation Exclusivity Author Rank. *Information Processing & Management* 52, 3 (2016), 374–386.

[3] Tehmina Amjad, Ying Ding, Ali Daud, Jian Xu, and Vincent Malic. 2015. Topic-based heterogeneous rank. *Scientometrics* 104, 1 (2015), 313–334.

[4] Tehmina Amjad, Ying Ding, Jian Xu, Chenwei Zhang, Ali Daud, Jie Tang, and Min Song. 2017. Standing on the shoulders of giants. *Journal of Informetrics* 11, 1 (2017), 307–323.
[5] Xiaomei Bai, Ivan Lee, Zhaolong Ning, Amr Tolba, and Feng Xia. 2017. The Role of Positive and Negative Citations in Scientific Evaluation. *IEEE Access* 5, 99 (2017), 17607–17617.
[6] Lutz Bornmann and Robin Haunschild. 2016. How to normalize Twitter counts? A first attempt based on journals in the Twitter Index. *Scientometrics* 107, 3 (2016), 1405–1422.
[7] Xuanyu Cao, Yan Chen, and K. J. Ray Liu. 2016. A data analytic approach to quantifying scientific impact. *Journal of Informetrics* 10, 2 (2016), 471–484.
[8] Aaron Clauset, Daniel B Larremore, and Roberta Sinatra. 2017. Data-driven predictions in the science of science. *Science* 355, 6324 (2017), 477–480.
[9] Rodrigo Costas, Zohreh Zahedi, and Paul Wouters. 2015. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science & Technology* 66, 10 (2015), 2003–2019.
[10] Yuxiao Dong, Reid Johnson, and Nitesh Chawla. 2016. Can Scientific Impact Be Predicted? *IEEE Transactions on Big Data* 2, 1 (2016), 18–30.
[11] Marcel Dunaiski, Willem Visser, and Jaco Geldenhuys. 2016. Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics* 10, 2 (2016), 392–407.
[12] Leo Egghe. 2006. Theory and practise of the g-index. *Scientometrics* 69, 1 (2006), 131–152.
[13] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, and B. Uzzi. 2018. Science of science. *Science* 359, 6379 (2018).
[14] Eugene Garfield. 2006. The history and meaning of the journal impact factor. *Jama* 295, 1 (2006), 90–93.
[15] Raphael H. Heiberger and Oliver J. Wieczorek. 2016. Choosing Collaboration Partners. How Scientific Success in Physics Depends on Network Positions. *arXiv:1608.03251* (2016).
[16] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences of the United States of America* 102, 46 (2005), 16569–16572.
[17] Jon M. Kleinberg. 1999. Authoritative Sources in a Hyperlinked Environment. *J. ACM* 46, 5 (Sept. 1999), 604–632.
[18] Peter Klimek, Aleksandar S. Jovanovic, Rainer Egloff, and Reto Schneider. 2016. Successful fish go with the flow: citation impact prediction based on centrality measures for termCdocument networks. *Scientometrics* 107, 3 (2016), 1265–1282.
[19] Xiangjie Kong, Yajie Shi, Wei Wang, Kai Ma, Liangtian Wan, and Feng Xia. 2019. The Evolution of Turing Award Collaboration Network: Bibliometric-level and Network-level Metrics. *IEEE Transactions on Computational Social Systems* (2019). https://doi.org/10.1109/TCSS.2019.2950445
[20] Thuc Duy Le, Lin Liu, Anna Tsykin, Gregory J Goodall, Bing Liu, Bing-Yu Sun, and Jiuyong Li. 2013. Inferring microRNA–mRNA causal regulatory relationships from expression data. *Bioinformatics* 29, 6 (2013), 765–771.
[21] Jiuyong Li, Thuc Duy Le, Lin Liu, Jixue Liu, Zhou Jin, Bingyu Sun, and Saisai Ma. 2016. From observational studies to causal rule mining. *ACM Transactions on Intelligent Systems and Technology (TIST)* 7, 2 (2016), 14.
[22] Liangyue Li and Hanghang Tong. 2015. The Child is Father of the Man:Foresee the Success at the Early Stage. In *Acm Sigkdd International Conference on Knowledge Discovery & Data Mining*. 655–664.
[23] Ronghua Liang and Xiaorui Jiang. 2016. Scientific ranking over heterogeneous academic hypernetwork. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI, 20–26.
[24] Jiaying Liu, Feng Xia, Lei Wang, Bo Xu, Xiangjie Kong, Hanghang Tong, and Irwin King. 2019. Shifu2: A Network Representation Learning Based Model for Advisor-advisee Relationship Mining. *IEEE Transactions on Knowledge and Data Engineering* (2019). https://doi.org/10.1109/TKDE.2019.2946825
[25] Lu Liu, Yang Wang, Roberta Sinatra, C. Lee Giles, Chaoming Song, and Dashun Wang. 2018. Hot streaks in artistic, cultural, and scientific careers. *Nature* 559, 7714 (2018), 396–399. https://doi.org/10.1038/s41586-018-0315-8
[26] Ye Liu, Luming Zhang, Liqiang Nie, Yan Yan, and David S Rosenblum. 2016. Fortune Teller: Predicting Your Career Path. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, Vol. 2016. 201–207.
[27] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The PageRank citation ranking: bringing order to the web. *Stanford InfoLab* (1999).
[28] Ana Severiano, João A Carriço, D Ashley Robinson, Mário Ramirez, and Francisco R Pinto. 2011. Evaluation of jackknife and bootstrap for defining confidence intervals for pairwise agreement measures.

PLoS One 6, 5 (2011), e19539.
[29] Richard M. Shiffrin. 2016. Drawing causal inference from Big Data. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7308–7309.
[30] Roberta Sinatra, Dashun Wang, Pierre Deville, Chaoming Song, and Albert-László Barabási. 2016. Quantifying the evolution of individual scientific impact. *Science* 354, 6312 (2016), aaf5239–aaf5239.
[31] Peter Spirtes and Kun Zhang. 2016. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics* 3, 1 (02 2016), 1–28.
[32] Clara Stegehuis, Nelly Litvak, and Ludo Waltman. 2015. Predicting the long-term citation impact of recent publications. *Journal of Informetrics* 9, 3 (2015), 642–657.
[33] Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*. AAAI.
[34] BA Van Houten, Jerry Phelps, Martha Barnes, and William A Suk. 2000. Evaluating scientific impact. *Environmental health perspectives* 108, 9 (2000), A392–A393.
[35] Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science* 342, 6154 (2013), 127–132.
[36] Senzhang Wang, Sihong Xie, Xiaoming Zhang, Zhoujun Li, Yueying He, and Yueying He. 2016. Coranking the Future Influence of Multiobjects in Bibliographic Network Through Mutual Reinforcement. *Acm Transactions on Intelligent Systems & Technology* 7, 4 (2016).
[37] Wei Wang, Shuo Yu, Teshome Megersa Bekele, Xiangjie Kong, and Feng Xia. 2017. Scientific collaboration patterns vary with scholars' academic ages. *Scientometrics* 112, 1 (2017), 329–343.
[38] Hao Wu, Maoyuan Sun, Peng Mi, Nikolaj Tatti, Chris North, and Naren Ramakrishnan. 2018. Interactive Discovery of Coordinated Relationship Chains with Maximum Entropy Models. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 1 (2018), 7.
[39] Feng Xia, Xiaoyan Su, Wang Wei, Chenxin Zhang, Zhaolong Ning, and Ivan Lee. 2016. Bibliographic Analysis ofNatureBased on Twitter and Facebook Altmetrics Data. *Plos One* 11, 12 (2016).
[40] Feng Xia, Wei Wang, Teshome Megersa Bekele, and Huan Liu. 2017. Big Scholarly Data: A Survey. *IEEE Transactions on Big Data* 3, 1 (2017), 18–35.
[41] Yang Yang, Jie Tang, and Juanzi Li. 2018. Learning to Infer Competitive Relationships in Heterogeneous Networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 1 (2018), 12.
[42] Dejian Yu, Wanru Wang, Shuai Zhang, Wenyu Zhang, and Rongyu Liu. 2017. A multiple-link, mutually reinforced journal-ranking model to measure the prestige of journals. *Scientometrics* 111, 1 (2017), 521–542.
[43] Jun Zhang, Zhaolong Ning, Xiaomei Bai, Xiangjie Kong, Jinmeng Zhou, and Feng Xia. 2017. Exploring time factors in measuring the scientific impact of scholars. *Scientometrics* 112, 3 (2017), 1301–1321.